Science at the heart of medicine

Big data

Vagish Hemmige

Jan 29th, 2024

EINSTEIN

Albert Einstein College of Medicine

- Disclosures
  - None related to this talk

Albert Einstein College of Medicine

Aims

- Identify scholarly questions best answered via the use of large data sets
- Describe the benefits and drawbacks of the use of large data sets
- Appreciate the importance of assessing data validity with national data

Big data

- Increasingly available and used
  - Terabyte-sized data sets can now be analyzed using a desktop
  - This would have required significant resources a decade ago
  - Electronic medical records as mandated by the HITECH act

Benefits of big data

- Can answer questions that could not be answered any other way
  - Rare conditions or treatments
  - Rare patient populations
- External validity since large data sets are more representative of national populations
  - Greater diversity of patient populations
  - Greater diversity of treatments

EINSTEIN
Albert Einstein College of Medicine

However…

- Big data is not the solution to all problems!
  - Most of the time, the data were collected for a purpose different from the reason you want to use it for
  - One has to carefully assess the reason a database was created, and the detailed processes that went into the data-generating process
  - Statisticians can crunch the numbers—***you*** have to spot problems as statisticians will not always be able to interpret the data well enough to identify when results are nonsensical

EINSTEIN

Albert Einstein College of Medicine

Sources of big data

- Health care systems
- Billing entities (insurance companies)
- Federal government or allied entities
  - Medicare
  - UNOS (transplant)
- Cohort studies, many of which collect large numbers of variables and also may bank blood samples
- Personal devices/wearables
- Apps

EINSTEIN
Albert Einstein College of Medicine

Variables

- Demographics
- Claims
- Procedures
- Laboratory values (not typically present in insurance data)
- Other test results (again, not typically present in insurance data)
- Death (very well validated in federal data)

Claims data

- Billing claims can be a very powerful
  - Diagnosis codes (ICD 9/ICD 10)
  - Procedures (CPT)
- However, they can be very unwieldy to work with
  - 69,832 ICD 10 codes
  - Yearly changes

{W56.22xA}

Struck by orca, initial encounter

Ellery Addington-White – Digital Mixed Media
Ellery is the Student Director at Center for Entrepreneurship and Liberal Education at Beloit College.

{Z73.4}

Inadequate social skills, not elsewhere classified

Erika Samlowski – Colored Pencil 8" x 10"
Erika is a 2nd year medical student at the Medical College of Wisconsin. She is interested in pursuing a career in Plastic/Reconstructive Surgery.

https://www.icd10illustrated.com/products/book

EINSTEIN
Albert Einstein College of Medicine

Claims data

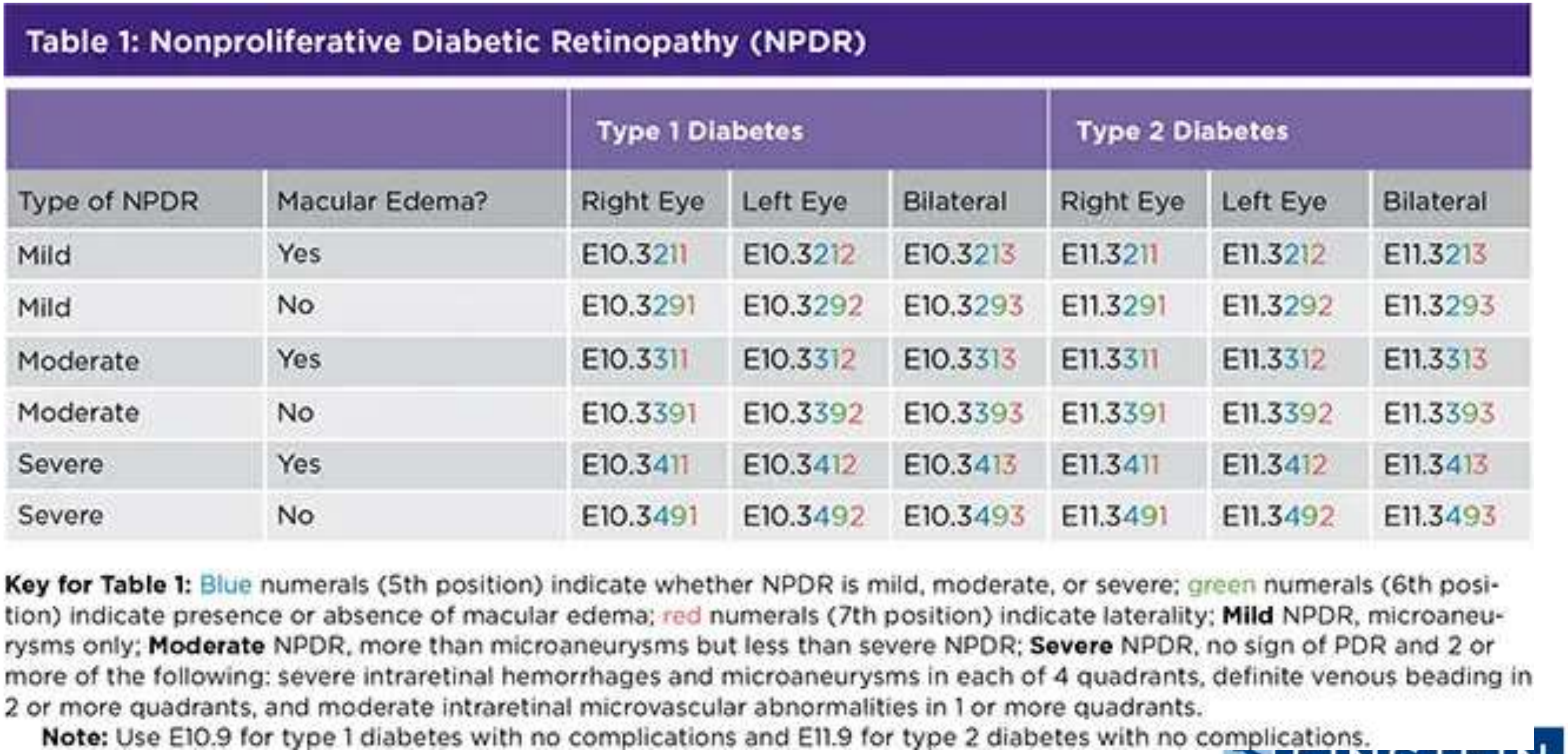


Table 1: Nonproliferative Diabetic Retinopathy (NPDR)

		Type 1 Diabetes			Type 2 Diabetes		
Type of NPDR	Macular Edema?	Right Eye	Left Eye	Bilateral	Right Eye	Left Eye	Bilateral
Mild	Yes	E10.3211	E10.3212	E10.3213	E11.3211	E11.3212	E11.3213
Mild	No	E10.3291	E10.3292	E10.3293	E11.3291	E11.3292	E11.3293
Moderate	Yes	E10.3311	E10.3312	E10.3313	E11.3311	E11.3312	E11.3313
Moderate	No	E10.3391	E10.3392	E10.3393	E11.3391	E11.3392	E11.3393
Severe	Yes	E10.3411	E10.3412	E10.3413	E11.3411	E11.3412	E11.3413
Severe	No	E10.3491	E10.3492	E10.3493	E11.3491	E11.3492	E11.3493

Key for Table 1: Blue numerals (5th position) indicate whether NPDR is mild, moderate, or severe; green numerals (6th position) indicate presence or absence of macular edema; red numerals (7th position) indicate laterality; **Mild** NPDR, microaneurysms only; **Moderate** NPDR, more than microaneurysms but less than severe NPDR; **Severe** NPDR, no sign of PDR and 2 or more of the following: severe intraretinal hemorrhages and microaneurysms in each of 4 quadrants, definite venous beading in 2 or more quadrants, and moderate intraretinal microvascular abnormalities in 1 or more quadrants.

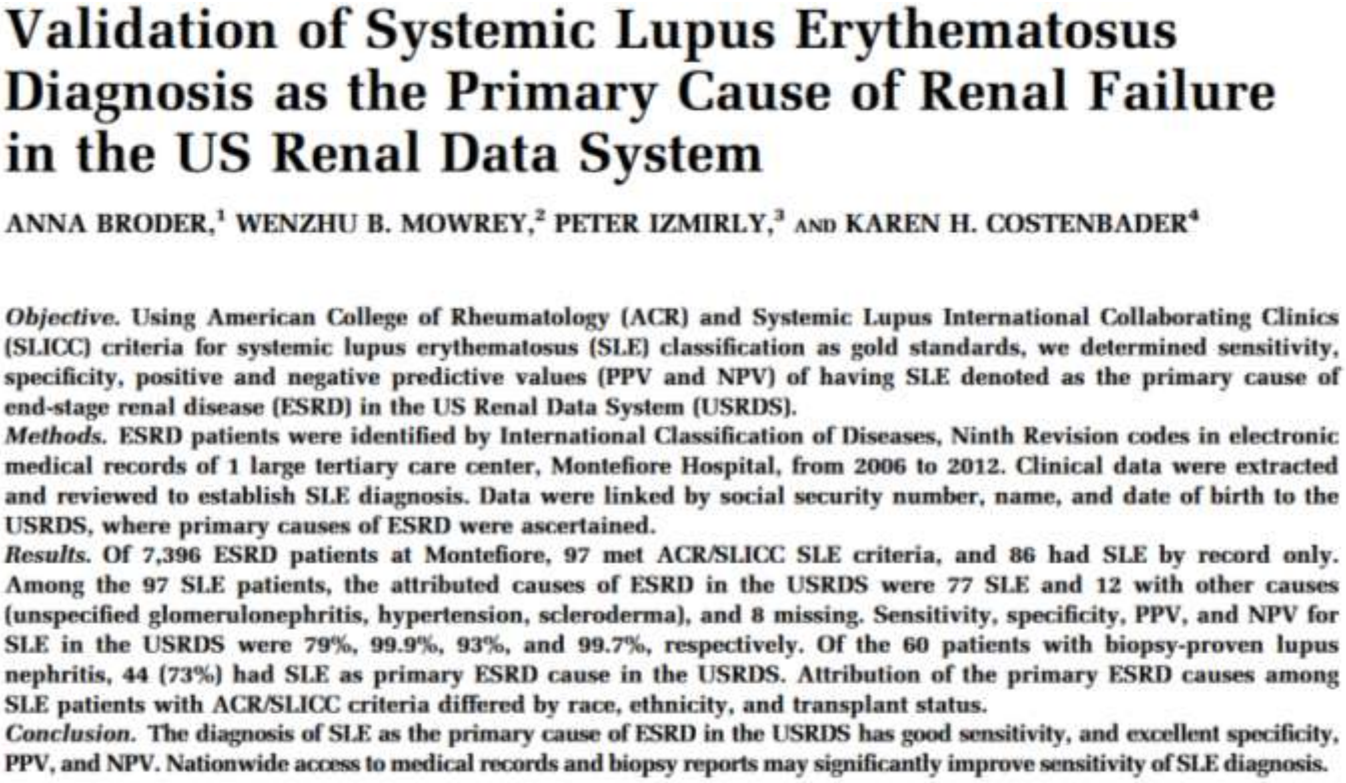
Note: Use E10.9 for type 1 diabetes with no complications and E11.9 for type 2 diabetes with no complications.

https://www.aao.org/eyenet/article/new-icd-10-codes-diabetic-retinopathy-amd

Challenges with claims data

- False positives
  - A physician sees a patient with onset swollen legs and bills that day under a heart failure diagnosis
  - However, workup shows a normal echocardiogram and abnormal creatinine with high degree of proteinuria, so ultimately the correct diagnosis of nephrotic syndrome is made. But the prior code is NOT changed retroactively
  - Solution: Many will only use a claim as a diagnosis if at least two outpatient visits, or one inpatient visit, use the claim
- False negatives
  - Not much you can do about this
- Validity
  - Do claims for a diagnosis correlate to having it?

Prior literature—are claims diagnoses validated?



Validation of Systemic Lupus Erythematosus Diagnosis as the Primary Cause of Renal Failure in the US Renal Data System

ANNA BRODER,[1] WENZHU B. MOWREY,[2] PETER IZMIRLY,[3] AND KAREN H. COSTENBADER[4]

Objective. Using American College of Rheumatology (ACR) and Systemic Lupus International Collaborating Clinics (SLICC) criteria for systemic lupus erythematosus (SLE) classification as gold standards, we determined sensitivity, specificity, positive and negative predictive values (PPV and NPV) of having SLE denoted as the primary cause of end-stage renal disease (ESRD) in the US Renal Data System (USRDS).
Methods. ESRD patients were identified by International Classification of Diseases, Ninth Revision codes in electronic medical records of 1 large tertiary care center, Montefiore Hospital, from 2006 to 2012. Clinical data were extracted and reviewed to establish SLE diagnosis. Data were linked by social security number, name, and date of birth to the USRDS, where primary causes of ESRD were ascertained.
Results. Of 7,396 ESRD patients at Montefiore, 97 met ACR/SLICC SLE criteria, and 86 had SLE by record only. Among the 97 SLE patients, the attributed causes of ESRD in the USRDS were 77 SLE and 12 with other causes (unspecified glomerulonephritis, hypertension, scleroderma), and 8 missing. Sensitivity, specificity, PPV, and NPV for SLE in the USRDS were 79%, 99.9%, 93%, and 99.7%, respectively. Of the 60 patients with biopsy-proven lupus nephritis, 44 (73%) had SLE as primary ESRD cause in the USRDS. Attribution of the primary ESRD causes among SLE patients with ACR/SLICC criteria differed by race, ethnicity, and transplant status.
Conclusion. The diagnosis of SLE as the primary cause of ESRD in the USRDS has good sensitivity, and excellent specificity, PPV, and NPV. Nationwide access to medical records and biopsy reports may significantly improve sensitivity of SLE diagnosis.

Arthritis Care Res (Hoboken). 2017 Apr;69(4):599-604.

Data validity

- Have to be very meticulous about looking for outliers. Yet it can be challenging to know what to do with them.
  - A patient in a database with A1C of 118% is clearly a typo. But is an A1c of 18.1% percent a typo or just an example of *extremely* poorly controlled diabetes?
  - Units issues
- Helps to know if there is a process for validating data
- Missing data also very frequently a problem
- Data shift

EINSTEIN
Albert Einstein College of Medicine

Data shift

- Best explained by examples
  - Before 2018, the data set stores death as a single variable, “Death”. Now it’s stored as two variables, “Social security-verified death” and “Clinic-reported death”
  - The way the data are stored in the variable changes
  - Behind the scenes changes in the data process
    - Increase in complications in patients with diabetes
    - Marked increase in HFpEF a few years ago

Data validity

- Death is a particularly challenging problem in non-federal data sets
  - Especially if patients are lost to follow-up
- A well-done national cohort data set will have a protocol for checking with Social Security or other sources to ascertain whether patients lost to follow up have died

Bottom line

- More data, more problems

- You frequently have to put in a *substantial* amount of time into learning your data set and its quirks

Data use agreement

- Technology has made it much easier to re-identify patients, even in de-identified data sets
- This is not theoretical—I once found out a former clinic patient of mine died while looking in a large national database and recognizing the patient
- Issues regarding data protection (no laptop, no flash drive, firewalls, etc.)
- Privacy—many datasets have rules such as no cells in a table smaller than a certain size, etc
- These rules are typically spelled out in a *data use agreement*, which is a binding legal contract

https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

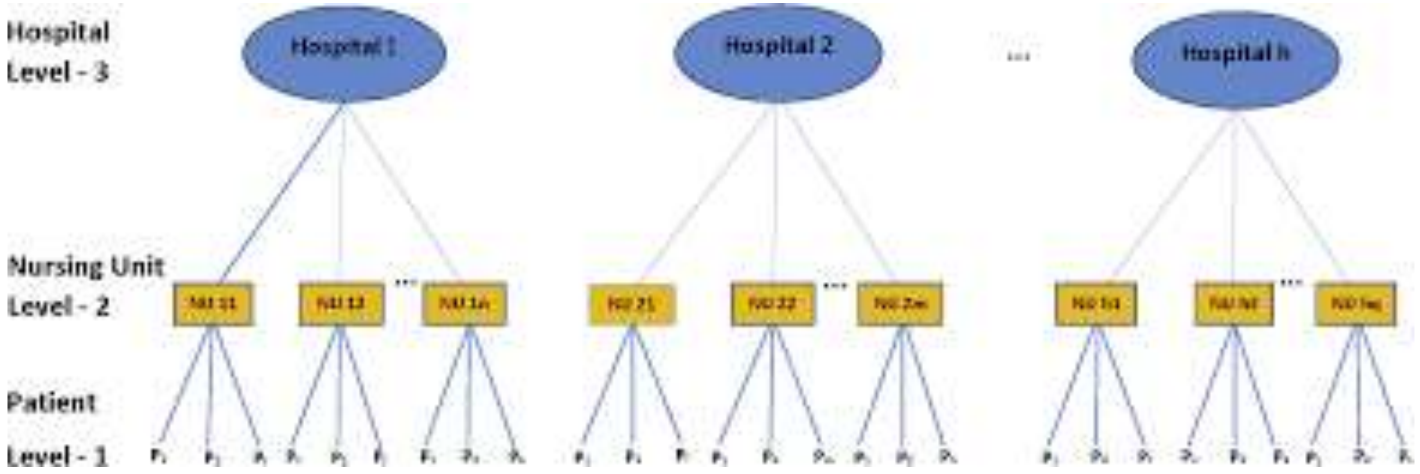
EINSTEIN
Albert Einstein College of Medicine

Study design issues

- Cross-sectional
- Cohort
- Case-control

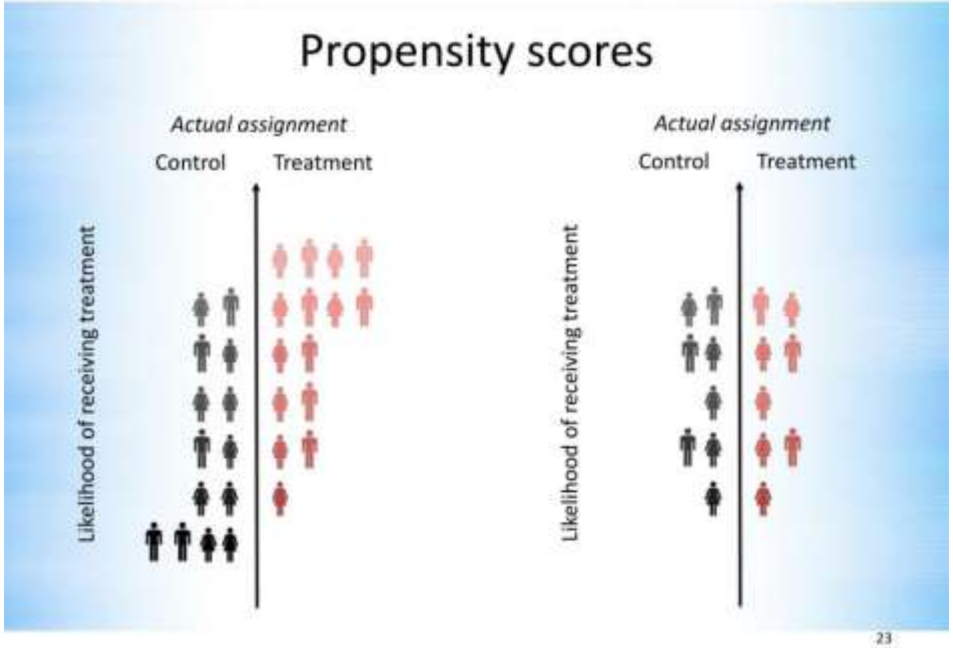
Statistical issues

- Standard statistical techniques may not be valid for data generated at multiple centers ("clustered data")
  - Multilevel modeling/mixed effects models
  - Marginal models



J Thorac Cardiovasc Surg. 2018 Jan;155(1):210-211.

- Other statistical techniques which may be appropriate:
  - > Matching (many algorithms exist, never as straightforward as one thinks)
  - > Propensity score
    - Calculate a model which predicts, based on baseline data, whether a patient received a treatment or not
    - Compare patients with similar scores who received different treatments



Propensity scores

https://www.linkedin.com/pulse/types-propensity-score-matching-jason-shafrin

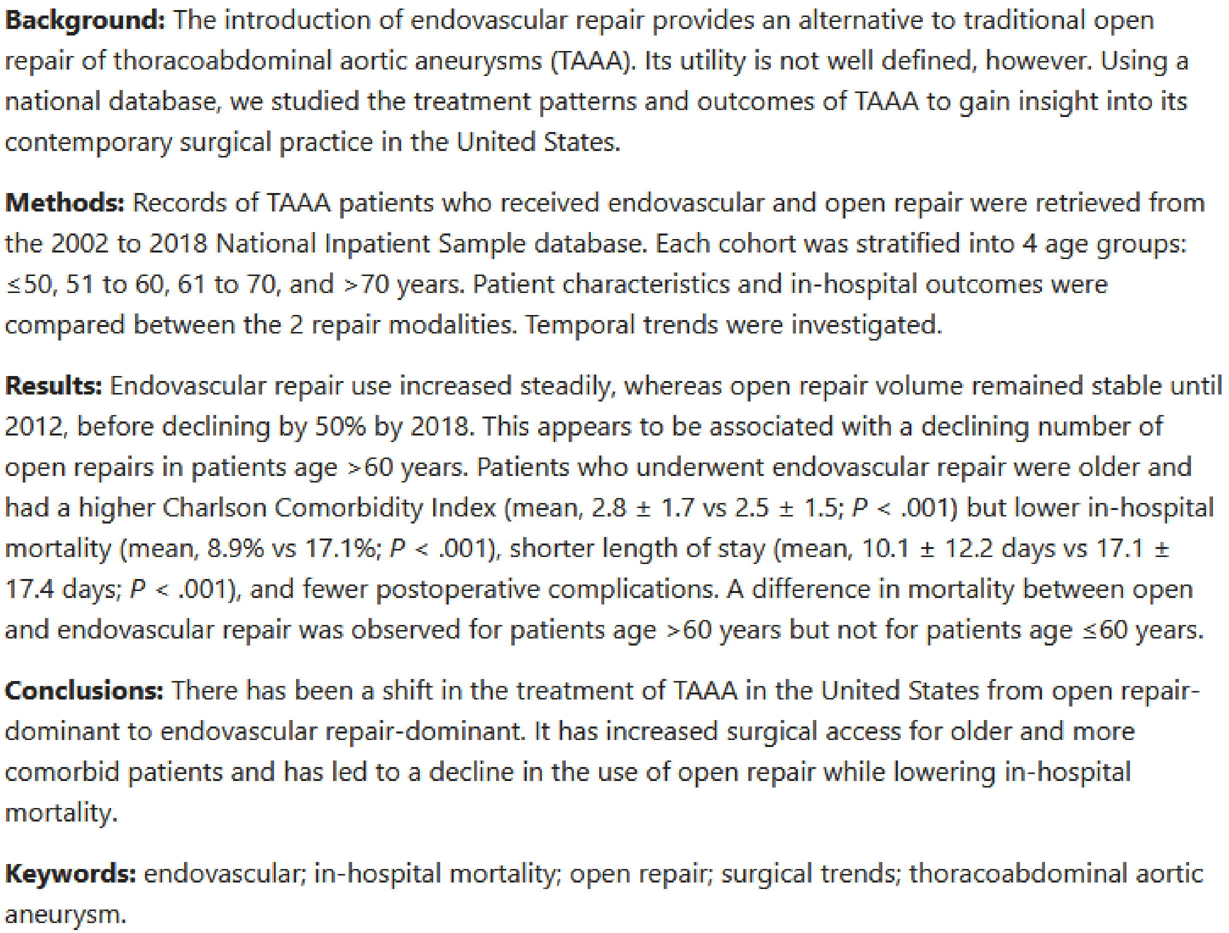
National inpatient sample

- 20% sample of all inpatient discharges
- Reasonable cost to acquire data
- https://hcup-us.ahrq.gov/nisoverview.jsp
- Cannot identify individual patients
- Variables:
  - Limited demographics
  - Billing codes
  - Procedural codes
  - Costs and length of stay
  - Outcome at discharge (alive or dead)

EINSTEIN

Albert Einstein College of Medicine

National inpatient sample



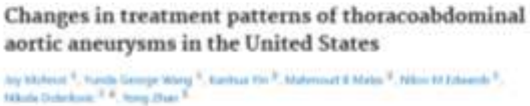
Background: The introduction of endovascular repair provides an alternative to traditional open repair of thoracoabdominal aortic aneurysms (TAAA). Its utility is not well defined, however. Using a national database, we studied the treatment patterns and outcomes of TAAA to gain insight into its contemporary surgical practice in the United States.

Methods: Records of TAAA patients who received endovascular and open repair were retrieved from the 2002 to 2018 National Inpatient Sample database. Each cohort was stratified into 4 age groups: ≤50, 51 to 60, 61 to 70, and >70 years. Patient characteristics and in-hospital outcomes were compared between the 2 repair modalities. Temporal trends were investigated.

Results: Endovascular repair use increased steadily, whereas open repair volume remained stable until 2012, before declining by 50% by 2018. This appears to be associated with a declining number of open repairs in patients age >60 years. Patients who underwent endovascular repair were older and had a higher Charlson Comorbidity Index (mean, 2.8 ± 1.7 vs 2.5 ± 1.5; *P* < .001) but lower in-hospital mortality (mean, 8.9% vs 17.1%; *P* < .001), shorter length of stay (mean, 10.1 ± 12.2 days vs 17.1 ± 17.4 days; *P* < .001), and fewer postoperative complications. A difference in mortality between open and endovascular repair was observed for patients age >60 years but not for patients age ≤60 years.

Conclusions: There has been a shift in the treatment of TAAA in the United States from open repair-dominant to endovascular repair-dominant. It has increased surgical access for older and more comorbid patients and has led to a decline in the use of open repair while lowering in-hospital mortality.

Keywords: endovascular; in-hospital mortality; open repair; surgical trends; thoracoabdominal aortic aneurysm.



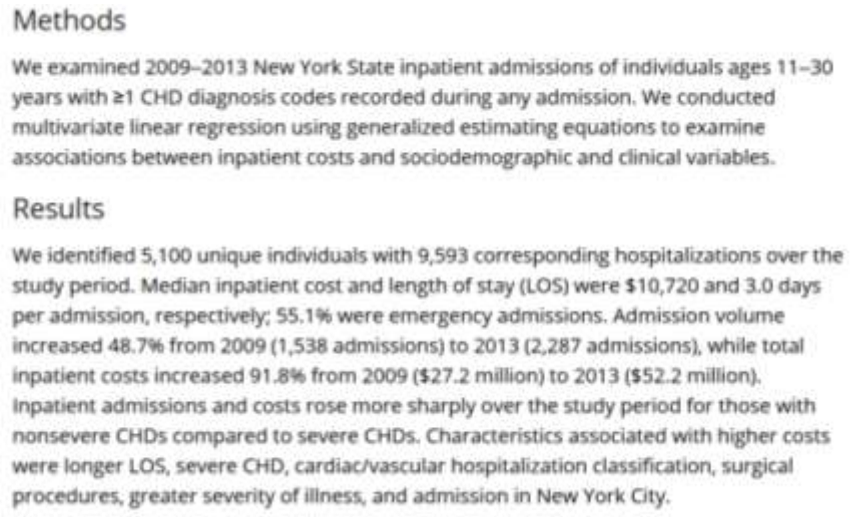
Changes in treatment patterns of thoracoabdominal aortic aneurysms in the United States

[illegible]

Albert Einstein College of Medicine

SPARCS

- Similar database for New York State only
- Also includes some outpatient data



Methods

We examined 2009–2013 New York State inpatient admissions of individuals ages 11–30 years with ≥1 CHD diagnosis codes recorded during any admission. We conducted multivariate linear regression using generalized estimating equations to examine associations between inpatient costs and sociodemographic and clinical variables.

Results

We identified 5,100 unique individuals with 9,593 corresponding hospitalizations over the study period. Median inpatient cost and length of stay (LOS) were $10,720 and 3.0 days per admission, respectively; 55.1% were emergency admissions. Admission volume increased 48.7% from 2009 (1,538 admissions) to 2013 (2,287 admissions), while total inpatient costs increased 91.8% from 2009 ($27.2 million) to 2013 ($52.2 million). Inpatient admissions and costs rose more sharply over the study period for those with nonsevere CHDs compared to severe CHDs. Characteristics associated with higher costs were longer LOS, severe CHD, cardiac/vascular hospitalization classification, surgical procedures, greater severity of illness, and admission in New York City.

https://www.health.ny.gov/statistics/sparcs/
https://onlinelibrary.wiley.com/doi/abs/10.1002/bdr2.1809

United States Renal Data System

- Links dialysis data for all ESRD patients in the United States with their Medicare claims data
- 3+ million patients in the system
- Very granular (ZIP code of residence, exact dates, and other very granular data
- I've merged with census data and geolocation data of dialysis centers or transplant centers to create an even richer database
- *Very long timelines*-can take months to years to get a project approved

JAMA Netw Open. 2024 Jan 2;7(1):e2350009.

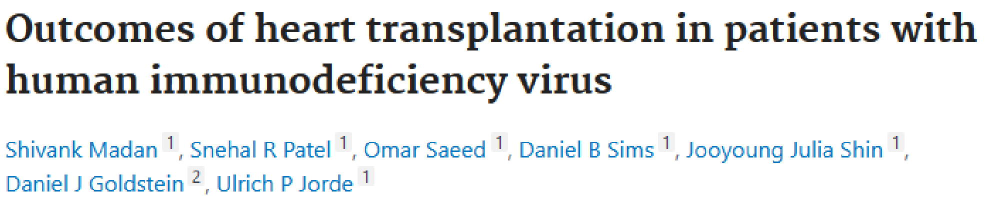
EINSTEIN
Albert Einstein College of Medicine

UNOS/SRTR

- The solid organ transplant network in the United States is run by an organization called UNOS
- They manage the waitlist and store waitlist data
- Transplant centers have mandatory submission of post-transplant outcomes data at specified intervals for each patient
- Issues with missingness and validation

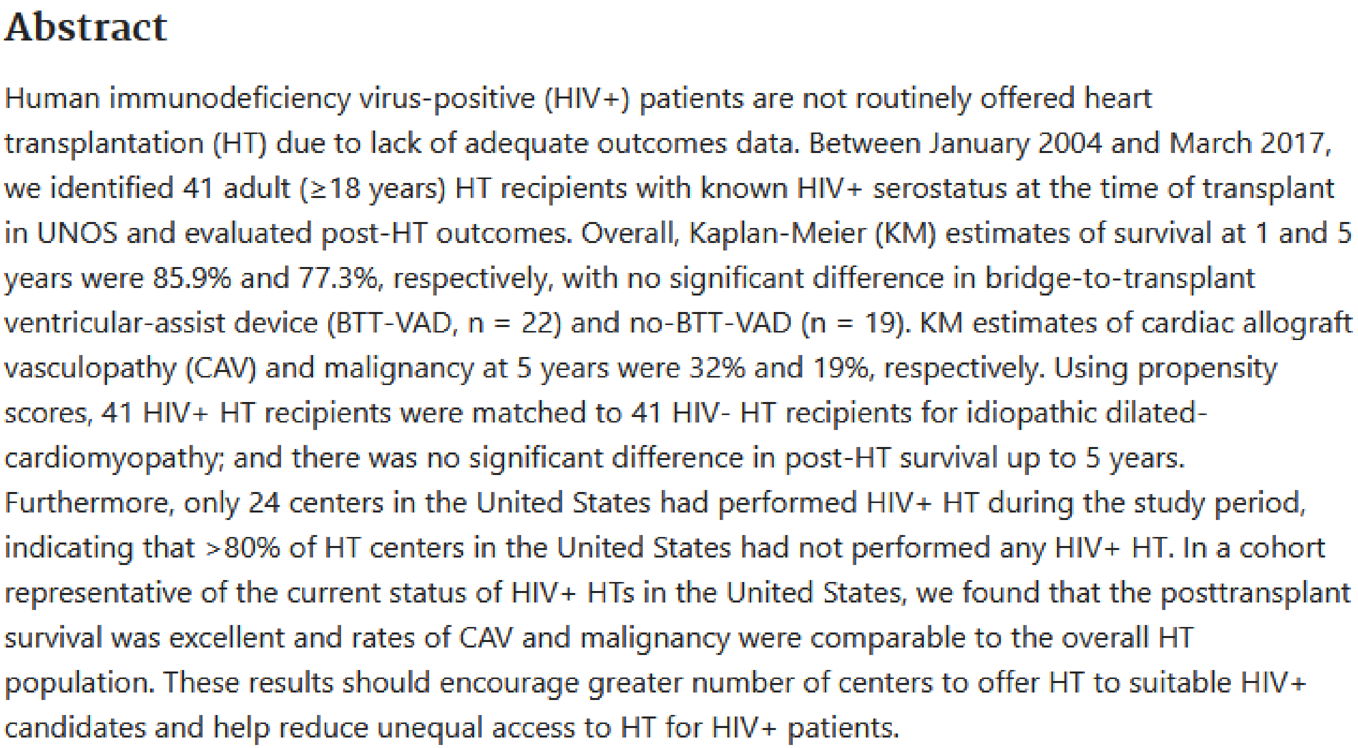
EINSTEIN
Albert Einstein College of Medicine

UNOS/SRTR



Outcomes of heart transplantation in patients with human immunodeficiency virus

Shivank Madan [1], Snehal R Patel [1], Omar Saeed [1], Daniel B Sims [1], Jooyoung Julia Shin [1], Daniel J Goldstein [2], Ulrich P Jorde [1]



Abstract

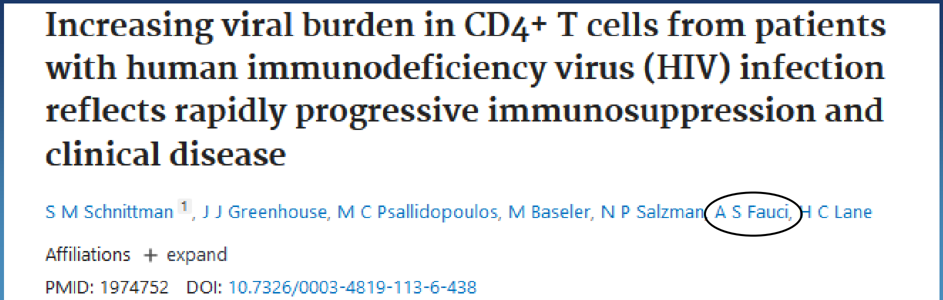
Human immunodeficiency virus-positive (HIV+) patients are not routinely offered heart transplantation (HT) due to lack of adequate outcomes data. Between January 2004 and March 2017, we identified 41 adult (≥18 years) HT recipients with known HIV+ serostatus at the time of transplant in UNOS and evaluated post-HT outcomes. Overall, Kaplan-Meier (KM) estimates of survival at 1 and 5 years were 85.9% and 77.3%, respectively, with no significant difference in bridge-to-transplant ventricular-assist device (BTT-VAD, n = 22) and no-BTT-VAD (n = 19). KM estimates of cardiac allograft vasculopathy (CAV) and malignancy at 5 years were 32% and 19%, respectively. Using propensity scores, 41 HIV+ HT recipients were matched to 41 HIV- HT recipients for idiopathic dilated-cardiomyopathy; and there was no significant difference in post-HT survival up to 5 years. Furthermore, only 24 centers in the United States had performed HIV+ HT during the study period, indicating that >80% of HT centers in the United States had not performed any HIV+ HT. In a cohort representative of the current status of HIV+ HTs in the United States, we found that the posttransplant survival was excellent and rates of CAV and malignancy were comparable to the overall HT population. These results should encourage greater number of centers to offer HT to suitable HIV+ candidates and help reduce unequal access to HT for HIV+ patients.

WIHS/MACS

- Multicenter AIDS Cohort Study
  - Started in 1984 (predates existence of the HIV test!)
  - Focused on MSM
  - Semiannual structured visits
  - https://www.niaid.nih.gov/research/multicenter-aids-cohort-study-public-data-set
- Women’s Interagency Health Study
  - Started in 1993
  - Cisgender women with HIV or “at risk for HIV”
  - Semiannual structured visits
  - Einstein was an original site and remains one today

EINSTEIN
Albert Einstein College of Medicine

MACS

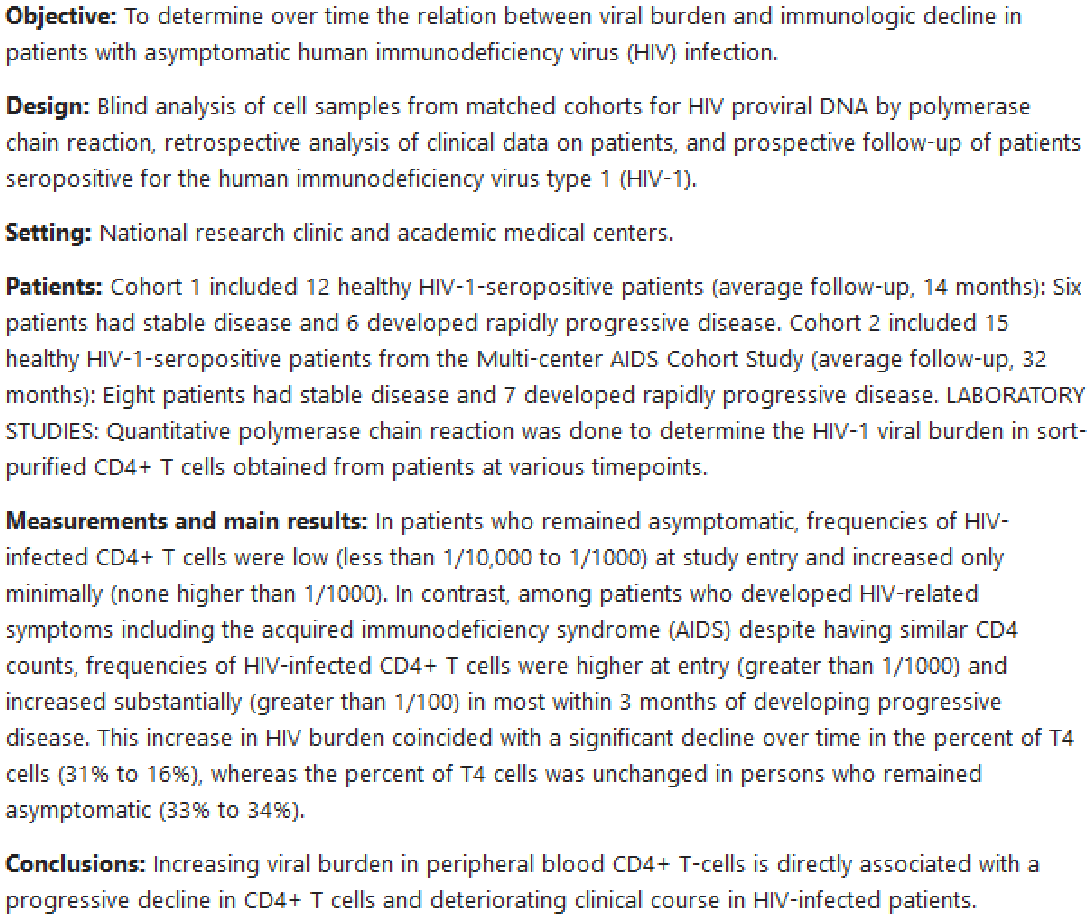


Increasing viral burden in CD4+ T cells from patients with human immunodeficiency virus (HIV) infection reflects rapidly progressive immunosuppression and clinical disease

S M Schnittman[1], J J Greenhouse, M C Psallidopoulos, M Baseler, N P Salzman, A S Fauci, H C Lane

Affiliations + expand

PMID: 1974752 DOI: 10.7326/0003-4819-113-6-438



Objective: To determine over time the relation between viral burden and immunologic decline in patients with asymptomatic human immunodeficiency virus (HIV) infection.

Design: Blind analysis of cell samples from matched cohorts for HIV proviral DNA by polymerase chain reaction, retrospective analysis of clinical data on patients, and prospective follow-up of patients seropositive for the human immunodeficiency virus type 1 (HIV-1).

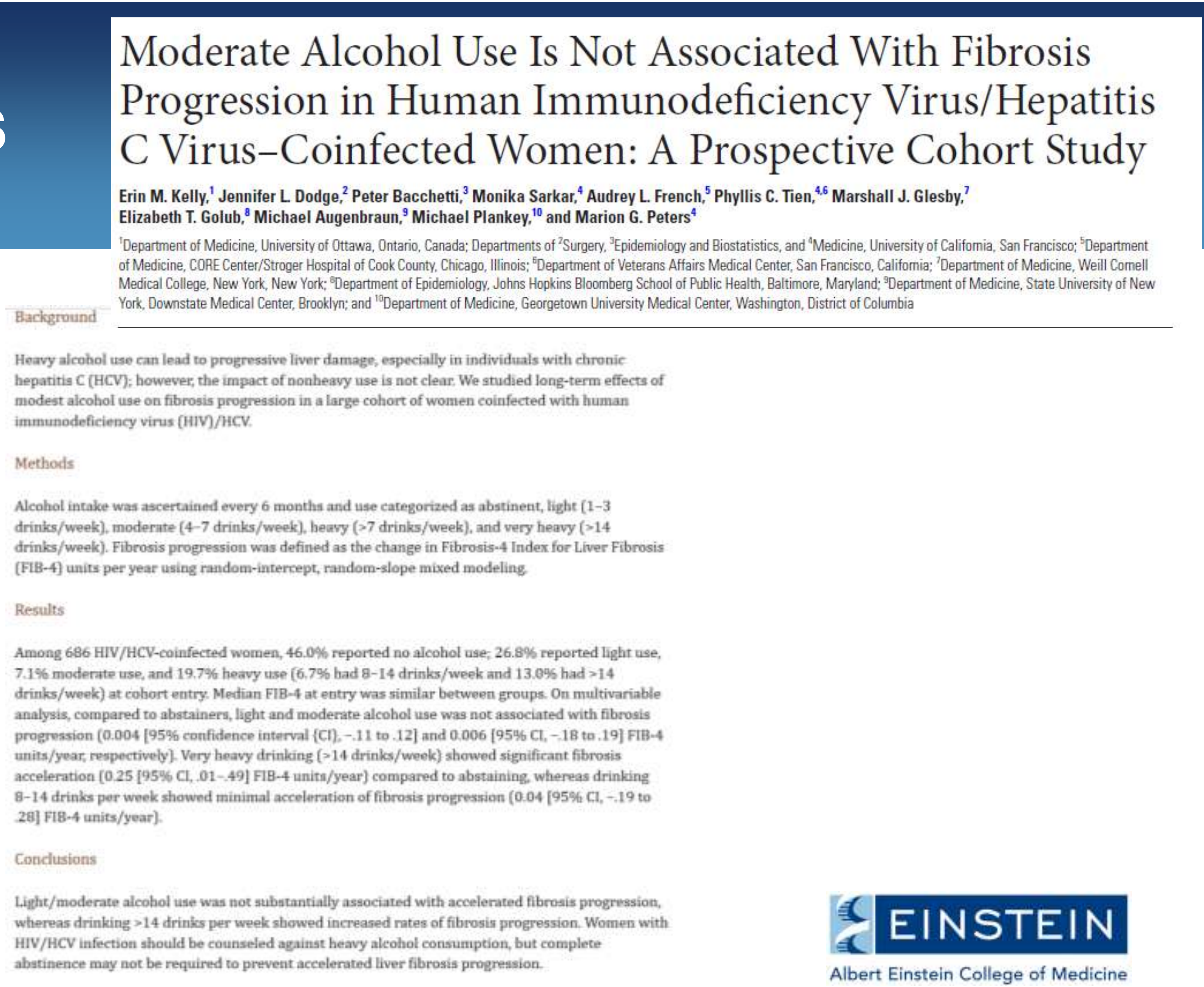
Setting: National research clinic and academic medical centers.

Patients: Cohort 1 included 12 healthy HIV-1-seropositive patients (average follow-up, 14 months): Six patients had stable disease and 6 developed rapidly progressive disease. Cohort 2 included 15 healthy HIV-1-seropositive patients from the Multi-center AIDS Cohort Study (average follow-up, 32 months): Eight patients had stable disease and 7 developed rapidly progressive disease. LABORATORY STUDIES: Quantitative polymerase chain reaction was done to determine the HIV-1 viral burden in sort-purified CD4+ T cells obtained from patients at various timepoints.

Measurements and main results: In patients who remained asymptomatic, frequencies of HIV-infected CD4+ T cells were low (less than 1/10,000 to 1/1000) at study entry and increased only minimally (none higher than 1/1000). In contrast, among patients who developed HIV-related symptoms including the acquired immunodeficiency syndrome (AIDS) despite having similar CD4 counts, frequencies of HIV-infected CD4+ T cells were higher at entry (greater than 1/1000) and increased substantially (greater than 1/100) in most within 3 months of developing progressive disease. This increase in HIV burden coincided with a significant decline over time in the percent of T4 cells (31% to 16%), whereas the percent of T4 cells was unchanged in persons who remained asymptomatic (33% to 34%).

Conclusions: Increasing viral burden in peripheral blood CD4+ T-cells is directly associated with a progressive decline in CD4+ T cells and deteriorating clinical course in HIV-infected patients.

WIHS



Moderate Alcohol Use Is Not Associated With Fibrosis Progression in Human Immunodeficiency Virus/Hepatitis C Virus–Coinfected Women: A Prospective Cohort Study

Erin M. Kelly,[1] Jennifer L. Dodge,[2] Peter Bacchetti,[3] Monika Sarkar,[4] Audrey L. French,[5] Phyllis C. Tien,[4,6] Marshall J. Glesby,[7] Elizabeth T. Golub,[8] Michael Augenbraun,[9] Michael Plankey,[10] and Marion G. Peters[4]

[1]Department of Medicine, University of Ottawa, Ontario, Canada; Departments of [2]Surgery, [3]Epidemiology and Biostatistics, and [4]Medicine, University of California, San Francisco; [5]Department of Medicine, CORE Center/Stroger Hospital of Cook County, Chicago, Illinois; [6]Department of Veterans Affairs Medical Center, San Francisco, California; [7]Department of Medicine, Weill Cornell Medical College, New York, New York; [8]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland; [9]Department of Medicine, State University of New York, Downstate Medical Center, Brooklyn; and [10]Department of Medicine, Georgetown University Medical Center, Washington, District of Columbia

Background

Heavy alcohol use can lead to progressive liver damage, especially in individuals with chronic hepatitis C (HCV); however, the impact of nonheavy use is not clear. We studied long-term effects of modest alcohol use on fibrosis progression in a large cohort of women coinfected with human immunodeficiency virus (HIV)/HCV.

Methods

Alcohol intake was ascertained every 6 months and use categorized as abstinent, light (1–3 drinks/week), moderate (4–7 drinks/week), heavy (>7 drinks/week), and very heavy (>14 drinks/week). Fibrosis progression was defined as the change in Fibrosis-4 Index for Liver Fibrosis (FIB-4) units per year using random-intercept, random-slope mixed modeling.

Results

Among 686 HIV/HCV-coinfected women, 46.0% reported no alcohol use; 26.8% reported light use, 7.1% moderate use, and 19.7% heavy use (6.7% had 8–14 drinks/week and 13.0% had >14 drinks/week) at cohort entry. Median FIB-4 at entry was similar between groups. On multivariable analysis, compared to abstainers, light and moderate alcohol use was not associated with fibrosis progression (0.004 [95% confidence interval {CI}, −.11 to .12] and 0.006 [95% CI, −.18 to .19] FIB-4 units/year, respectively). Very heavy drinking (>14 drinks/week) showed significant fibrosis acceleration (0.25 [95% CI, .01–.49] FIB-4 units/year) compared to abstaining, whereas drinking 8–14 drinks per week showed minimal acceleration of fibrosis progression (0.04 [95% CI, −.19 to .28] FIB-4 units/year).

Conclusions

Light/moderate alcohol use was not substantially associated with accelerated fibrosis progression, whereas drinking >14 drinks per week showed increased rates of fibrosis progression. Women with HIV/HCV infection should be counseled against heavy alcohol consumption, but complete abstinence may not be required to prevent accelerated liver fibrosis progression.

EINSTEIN
Albert Einstein College of Medicine

Data sets that have been used for student projects in the past

Montefoire Einstein Cancer Cohort
Einstein Aging Study
Center for AIDS Research (CFAR) -- HIV Clinical Cohort Database (CCDB)
World Trade Center 9/11 Firefighter Cohort
Longevity Gene Project (LGP)
LonGenity
Hispanic Community Health Study/Study of Latinos (HCHS/SOL)
Women's Health Initiative (WHI)
Study of Women's Health Across the Nation (SWAN)
Mulitcenter AIDS Cohort Study - Women's Interagency HIV Study Combined Cohort Study (MACS/WIHS CCS)
Diabetes Prevention Program (DPP)
Diabetes Prevention Program Outcomes Study (DPPOS)
National Health and Nutrition Examination Survey (NHANES)
United Network for Organ Sharing (UNOS)
National Immunization Survey (NIS)
Healthcost and Utilization Project Kids Inpatient Database (HCUP-KID)
Healthcost and Utilization Project National Inpatient Sample (HCUP-NIS)
National Surgical Quality Improvement Program (NSQIP)
United States Renal Data System (USRDS)
Statewide Planning and Research Cooperative System (SPARCS)
National Health Interview Survey (NHIS)
All of Us Genomic Data
UK Biobank
Agricultural Health Study (AHS)
NIH-AARP Diet and Health Study
Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)

EINSTEIN
Albert Einstein College of Medicine